UCRL-TR-155600

Start/End Delays of Voiced and Unvoiced Speech Signals

Aaron Herrnstein
University of California Davis, Department of Applied Science & Lawrence Livermore National Laboratory

September 24, 1999

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

UCRL-TR-155600

Start/End Delays of Voiced and Unvoiced Speech Signals

Aaron Herrnstein

University of California Davis, Department of Applied Science
&
Lawrence Livermore National Laboratory

9/24/99

Abstract:

Recent experiments using low power EM-radar like sensors (e.g, GEMs) have demonstrated a new method for measuring vocal fold activity and the onset times of voiced speech, as vocal fold contact begins to take place. Similarly the end time of a voiced speech segment can be measured. Secondly it appears that in most normal uses of American English speech, unvoiced-speech segments directly precede or directly follow voiced-speech segments. For many applications, it is useful to know typical duration times of these unvoiced speech segments. A corpus, assembled earlier of spoken "Timit" words, phrases, and sentences and recorded using simultaneously measured acoustic and EM-sensor glottal signals, from 16 male speakers, was used for this study. By inspecting the onset (or end) of unvoiced speech, using the acoustic signal, and the onset (or end) of voiced speech using the EM sensor signal, the average duration times for unvoiced segments preceding onset of vocalization were found to be 300ms, and for following segments, 500ms. An unvoiced speech period is then defined in time, first by using the onset of the EM-sensed glottal signal, as the onset-time marker for the voiced speech segment and end marker for the unvoiced segment. Then, by subtracting 300ms from the onset time mark of voicing, the unvoiced speech segment start time is found. Similarly, the times for a following unvoiced speech segment can be found. While data of this nature have proven to be useful for work in our laboratory, a great deal of additional work remains to validate such data for use with general populations of users. These procedures have been useful for applying optimal processing algorithms over time segments of unvoiced, voiced, and non-speech acoustic signals. For example, these data appear to be of use in speaker validation, in vocoding, and in denoising algorithms.

Introduction:

Centimeter wave length radars have been fashioned into Glottal Electromagnetic Micropower Sensors (GEMS) [1, 2] which are the basis of the technology used for the Speech Technology project [3] at Lawrence Livermore National Laboratory. Such sensors are used to monitor the motion of a speaker's glottis. The human voice uses two types of excitations: glottal activated (voiced) and air turbulence (unvoiced). We are studying the timing between the two excitations. It was observed that unvoiced signals typically have start times before and end times later than voiced signals (see Figure 1). The goal of the research presented in this paper was to determine the range and the maximum delay times between voiced and unvoiced signals. This information will be useful for speech recognizers which use the GEMS technology. Such recognizers consider a speaker to begin speaking at the start of

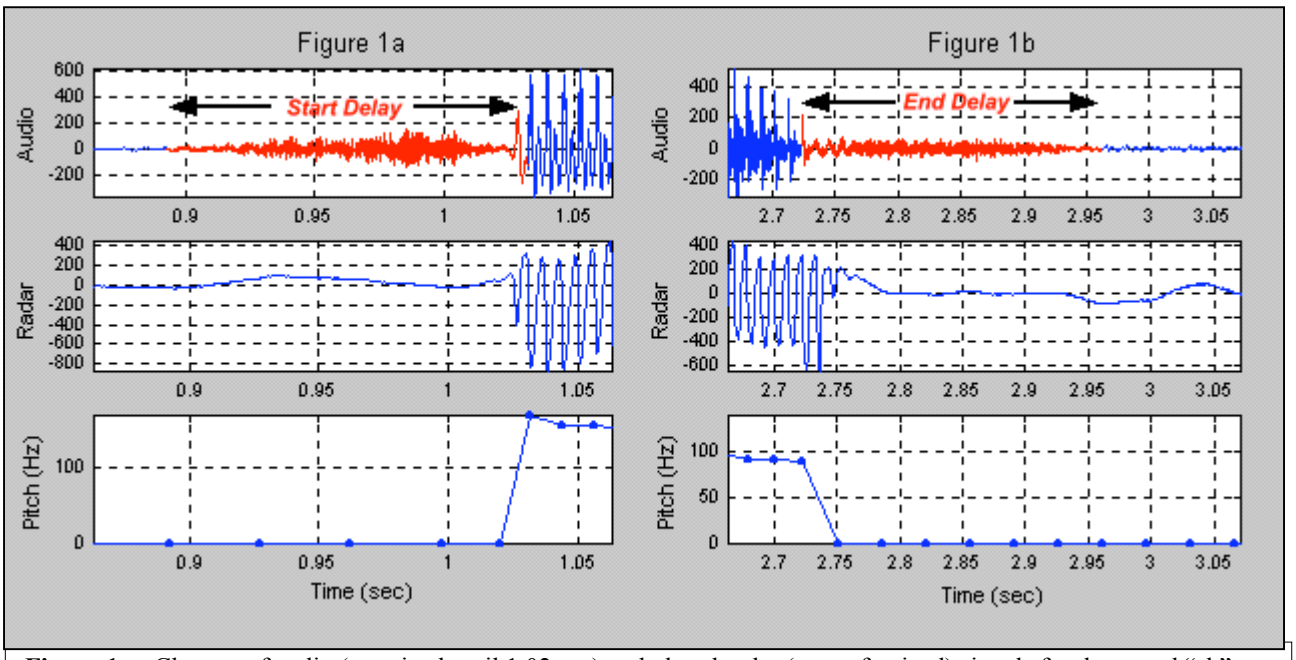

Figure 1a: Closeup of audio (unvoiced until 1.02 sec) and glottal radar (start of voiced) signals for the sound "sh" in the word "she." The resulting start delay (highlighted on graph) between unvoicing and voicing is 140 msec.

Figure 1b: The end of voiced and unvoiced signals of the sound "s" in the word "force." The resulting end delay is 230 msec. The pitch was calculated for both figures using an algorithum written by Burnett.

the voiced signal, and then terminate their speech with the end of the voiced signal. Finding the maximum delay times between voicing and unvoicing will allow data to be recorded of the previous and post unvoiced speech segments. This will provide a complete segment of speech containing audio and radar signals in their entirety.

Data was collected by Burnett and Gable of 15 males speaking 12 sentences [4]. Speakers pronounced each sentence 10 times producing 1800 data files for analysis [5]. Each file contained signals of the speaker's audio and glottal radar. These signals allowed calculations of the speaker's pitch using an algorithm written by Burnett [6]. The turn on/off times of the voicing onset were used as the start and end of the voiced signal. Unfortunately, no accurate method has yet to be found which will determine the start and end of the unvoiced signal reliably. These times were therefore chosen visually. The difference in the turn on times of the voiced and unvoiced signals was known as the start delay. Similarly, the end delay was the difference between the termination times of voiced and unvoiced signals. An example of the delay times is shown in Figure 1.

This method of collecting delay times turned out to have a few problems. First, the start of many unvoiced signals were not as well defined as in Figure 1. So a guess was often used. Secondly, choosing points visually became quite monotonous at times and unexpected human errors may have resulted. Perhaps the most significant problem was that a small portion of files only showed excessive pitch intensity in the middle of the signal (see Figure 2*a*) producing false turn on/off times for the voiced signal. Furthermore, the program used for data collection did not allow for a visual selection of these times. So for cases where the pitch algorithm gave poor results, the delay times were grossly over estimated in our first attempts.

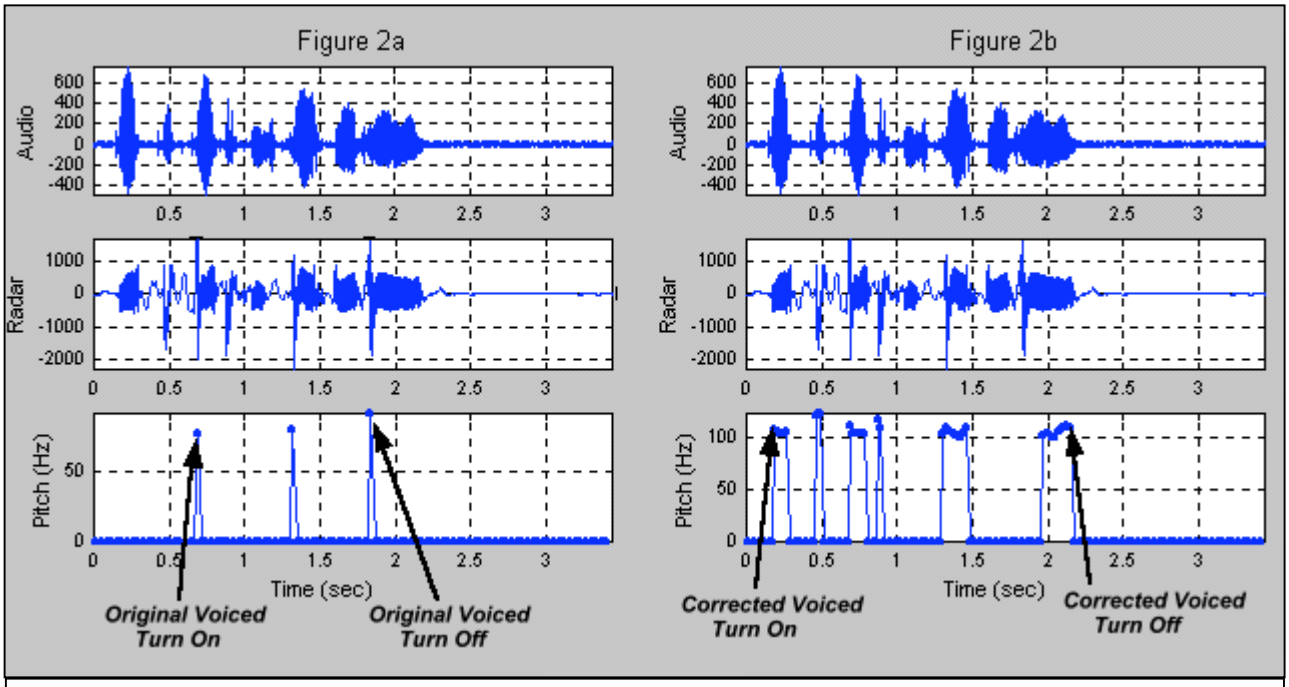

Figure 2a: Example of a file containing spikes in radar signal which produced inaccurate pitch calculations. As a result, incorrect turn on/off times of the voiced signal were recorded.

Figure 2b: Pitch results after applying a low pas filter. Correct voiced turn on/off times are now present.

The source of the pitch problem was noticed through inspection. It turned out that files which produced inaccurate pitch calculations contained one or more spikes in either the audio or radar signals. Once this correlation was made, a solution to the problem was developed. Originally, signals were filtered in the range of 60-650Hz before calculating the pitch. However, applying a 60-150Hz lowpass filter would essentially eliminate any spikes containing high frequencies from being used in such calculations. It turned out that each file required a slightly different lowpass filter to produce the proper turn on/off times of the voiced signal. A program was written that found a file's voiced start/end times for each 40Hz filter between 60-150Hz as well as the normal 60-650Hz filter. The earliest and latest times found by the program were used as the turn on/off times of the voiced signal (see Figure 2*b*). This program was applied to each file as a correction. As a result, error was reduced in the start/end times of the voiced signals, but the unvoiced signals were unaffected in that their times were still chosen visually.

Data Analysis and Results:

A small percentage of 1800 data files contained no delays (see Figure 3). Once the lowpass correction was applied, 20% of the files failed to show unvoiced precursors and only 0.5% contained no unvoicing after voiced turn offs. This suggests a very large probability of finding unvoiced speech at the end of a sentence. However, there is a small probability of no unvoiced occurrence at the beginning of a speaker's speech.

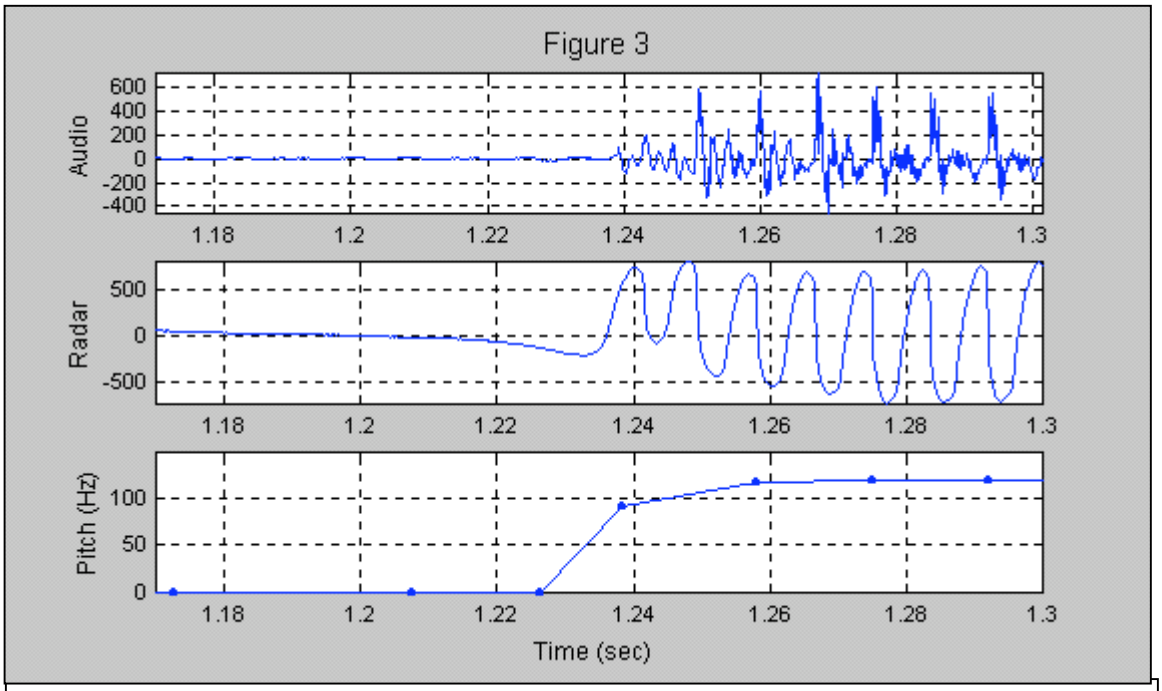

Figure 3: Example of a file containing no unvoicing signal.

An analysis was conducted on the corrected data (see Figure 4). Figures 4*a* and *b* show histograms of the delay times after the lowpass correction is applied. Figure 4*c* is a simple plot of the number of files contained within a given delay. Notice that both lines in the later plot level out at 1800 files (i.e., the total number of files used). It can be estimated from these graphs that the majority of unvoicing starts a maximum of (300 ± 20) msec ahead of voicing. Through the same inspection, most voiced signals terminate a maximum of (500 ± 20) msec before that of the unvoiced turn offs. Unvoiced signals typically fade into and out from noise at times near 20 msec. This time was therefore used as the error in the numbers listed above.

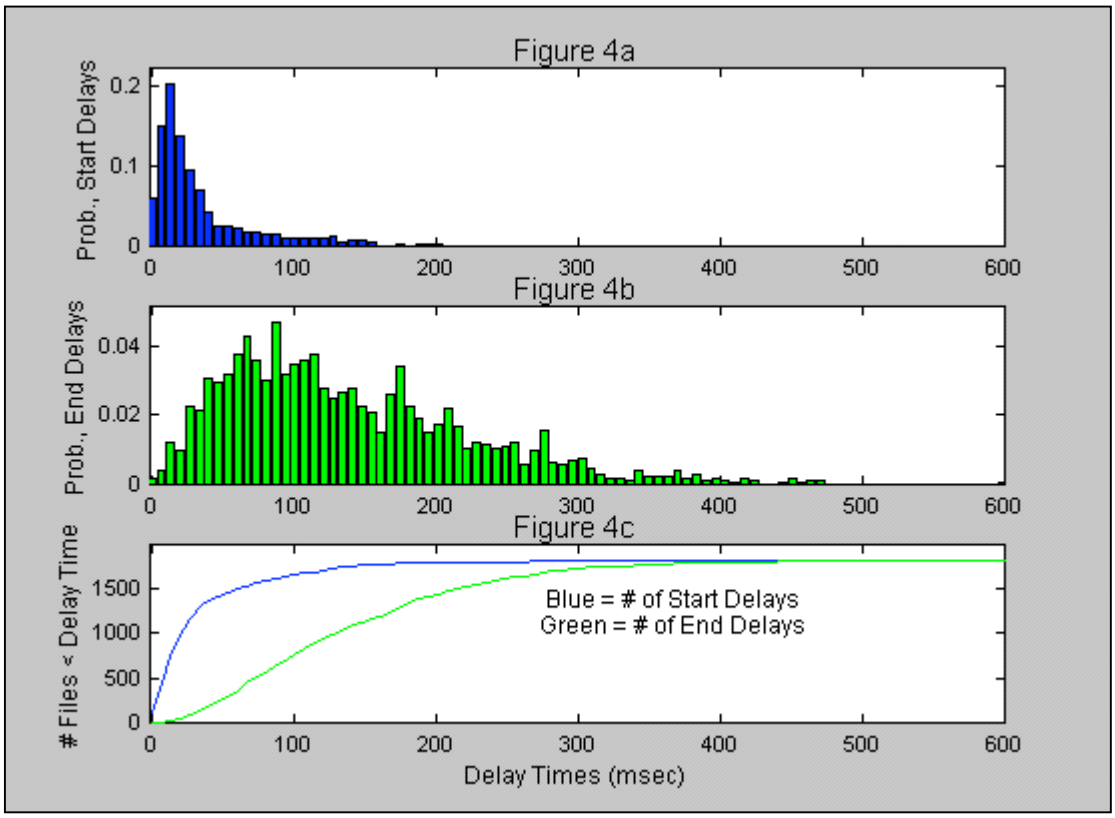

Figure 4a: Probability distribution of start time delays for corrected data.
Figure 4b: Probability distribution of end time delays for corrected data.
Figure 4c: Number of files with start/end delays less than a given delay time.

Conclusions:

The maximum start/end delays to be used in conjunction with a voice recognizer were found to be (300 ± 20) msec and (500 ± 20) msec. A lowpass filter correction is recommended to determine the turn on/off times of voiced speech. Using an algorithm like the one described under *Data Collection* will ignore any spikes in audio or radar signals. Such spikes in the GEMS data may have resulted from malfunctions in the radar circuitry. Though these errors may be corrected using a proper lowpass filter, modifications should be made to eliminate the possibility of their occurrence. Such errors may turn out to be crucial to future experimentation and analysis.

Acknowledgement:

I would like to thank John F. Holzrichter, Larry C. Ng, Greg C. Burnett, and Todd J. Gable for suggesting this project and for their assistance in carrying out the analyses in this paper.

References

[1] McEwan, T.E., work at Lawrence Livermore National Laboratory, U.S.Patents 5,345,471 and 5,361,070 (1994)

(2) Burnett, G. C. , "The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract," Ph.D. Thesis, Department of Applied Science, University of California at Davis, 1999

[3] Holzrichter, J.F., Burnett, G.C., Ng, L.C., and Lea, W.A., "Speech Articulator Measurements Using Low Power EM-Wave Sensors," J. Acoust. Soc. Am. 103(1), 1998, 622-625

[4] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, Massachusetts Institute of Technology, Stanford Research Institute, Texas Instruments, National Institute of Standards and Technology, http://morph.ldc.upenn.edu/

[5] Burnett, G. C., Gable, T.J., 8 CD set of GEMS data produced by the Speech Research Group of University of California at Davis and Lawrence Livermore National Laboratory, available as UCRL – MI – 132776, February 1, 1999

[6] Burnett, G.C., Gable, T.J., Ng, L.C., and Holzrichter, J.F. "Accurate, Inexpensive, and Noise-Immune Pitch Extraction Method Using Non-Acoustic Low Power Electromagnetic Sensors," 1998, submitted for publication, UCRL – JC – 130823

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551